Social B(eye)as: Human and Machine Descriptions of People Images

Pınar Barlas,[1] **Kyriakos Kyriakou,**[1] **Styliani Kleanthous,**[1,2] **& Jahna Otterbacher**[1,2]
[1]Research Centre on Interactive Media, Smart Systems and Emerging Technologies (Nicosia, CYPRUS)
[2]Cyprus Center for Algorithmic Transparency, Open University of Cyprus (Latsia, CYPRUS)
{pin.barlas, kyriakos093}@gmail.com, {styliani.kleanthous, jahna.otterbacher}@ouc.ac.cy

Abstract

Image analysis algorithms have become an indispensable tool in our information ecosystem, facilitating new forms of visual communication and information sharing. At the same time, they enable large-scale socio-technical research which would otherwise be difficult to carry out. However, their outputs may exhibit social bias, especially when analyzing people images. Since most algorithms are proprietary and opaque, we propose a method of auditing their outputs for social biases. To be able to compare how algorithms interpret a controlled set of people images, we collected descriptions across six image tagging algorithms. In order to compare these results to human behavior, we also collected descriptions on the same images from crowdworkers in two anglophone regions. The dataset we present consists of tags from these eight taggers, along with a typology of concepts, and a python script to calculate vector scores for each image and tagger. Using our methodology, researchers can see the behaviors of the image tagging algorithms and compare them to those of crowdworkers. Beyond computer vision auditing, the dataset of human- and machine-produced tags, the typology, and the vectorization method can be used to explore a range of research questions related to both algorithmic and human behaviors.

Introduction

Computer vision is widely recognized as one of the success stories of modern machine learning. Although once restricted to domains such as the military, security and surveillance, or medical imaging, applications of computer vision are now commonly used in consumer domains, from social media to e-commerce sites. In particular, image processing and analysis, in which the content of an input image is inferred, is now being used extensively in the modern information ecosystem. Users have become accustomed to the capabilities it facilitates, such as searching for and/or sharing image content in real time, without the need to provide descriptive metadata detailing the content. Similarly, in fields such as interactive marketing, image processing has become an essential tool that enables professionals to learn about and/or engage audiences via platforms that facilitate more visual forms of communication.

Within the research community as well, computer vision algorithms are enabling large-scale studies of Web and social media phenomena, which would otherwise not be possible via manual analysis. In a study conducted before image analysis application programming interfaces (APIs) were readily available, Hu and colleagues (Hu et al. 2014) implemented a published algorithm for inferring concepts in Instagram images, in order to examine the nature of the content shared by participants. More recently, proprietary APIs are being used extensively by researchers of socio-technical systems. For instance, (Deeb-Swihart et al. 2017) used Face++ in their study of selfies on Instagram, to characterize the types of selfies shared and by whom. In a similar vein, Liu and colleagues (Liu et al. 2016) applied image analysis to users' Twitter profile photos, to infer aspects of their personalities. Inspired by Google Flu Trends[1], Garimella and colleagues used Imagga to analyze images on Instagram to glean information about public health (Garimella, Alfayad, and Weber 2016). Finally, (Kocabey et al. 2018) used Face++ in inferring people's body mass index (BMI) from social media profile pictures, to study the relationship between popularity and weight. The above are but a few examples of the manner in which proprietary computer vision algorithms are facilitating the work of researchers in the community.

However, despite the innovation brought about by the success of image analysis technology, these algorithmic processes are not infallible. Unfortunately, there have already been many (public) incidents, and consequently scientific studies, on the ways that machine learning applications can produce socially harmful results. One of the most well-known examples of offensive and discriminatory outputs in image tagging was the 2015 Google Photos incident, where a Black software engineer's photo, depicting himself and a friend, was labeled with the tag "gorillas." [2] A recent study has found an increased error rate in gender classification for people with darker skin (as compared to lighter skin) and women (as compared to men), where the disparity between error rates can be more than 30% between light-skinned men and dark-skinned women (Buolamwini and Gebru 2018). Another study found that Black men were more likely to be tagged with a negative emotion than White men, when using

[1]https://www.google.org/flutrends/about/

[2]https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Face++ and Microsoft's Face API[3] (Rhue 2018).

These biases, while readily apparent in research studies, may be hard to recognize in places where the APIs are applied. Given the pressures of the commercial software industry, it is fair to assume that developers using a given algorithmic service will be primarily checking whether the outputs align with their goals for using it - and may be unaware that the results could be biased. As a result, the systematic differences in how various social groups are treated by an API can be carried "downstream," going on to affect everyone using the products underpinned by the technology, reinforcing the discriminatory practices in society. Furthermore, the use of proprietary image analysis algorithms is on the rise, as they are being rapidly commercialized in what Gartner has described as the "Algorithm Economy."[4]

Democratizing proprietary algorithms

The deployment of Cognitive Services (CogS), a key industry response to the new Algorithm Economy, has further fueled the uptake of new AI technologies such as image processing, by providing developers a convenient (typically via REST APIs) and economical means to incorporate these capabilities in their products and services. In fact, Microsoft has referred to the "democratiz[ation of] AI" through CogS.[5]

At the same time, the use of third-party tools such as CogS represents a liability for developers in light of new legislation and industry standards surrounding the protection of citizens' personal data as well as automated processes used on such data. For instance, the IEEE is developing a standard on Algorithmic Bias Considerations,[6] looking to provide a certification process through which developers can demonstrate their adherence to best practices surrounding the use of algorithmic processes. However, it may be difficult - if not impossible - for them to ensure that the providers of algorithmic processes that they use (i.e., CogS) do the same.

Furthermore, the EU's newly adopted General Data Protection Regulation (GDPR) affects the routine use of machine learning algorithms in a number of ways. Article 4[7] defines profiling as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain aspects relating to a natural person." The processing of *images depicting people*, commonly shared on social platforms, by a tagging algorithm appears to constitute such automated processing. Per GDPR, the developer must be positioned to provide a "meaningful explanation" of these processes to data subjects (i.e., consumers).

Image tagging algorithms, provided as CogS, are *opaque technologies* and it is not always possible to predict or explain the outputs of their analyses. Burrell (Burrell 2016) has described three types of opacity, and the taggers we study exhibit all three. All are *proprietary services*, thus, the providers do not disclose detailed information about their behaviors (e.g., the full set of tags used). Furthermore, there are *technical barriers* associated with their lack of transparency, as they are all based on deep learning methods. Finally, *technical literacy* is an issue as many small companies and/or individual developers may not be experts in machine learning and thus, not positioned to understand how the taggers work and also explain this to others (e.g., consumers whose images they have processed with the APIs).

Sandvig and colleagues (Sandvig et al. 2014) advocate the auditing of algorithmic processes "from the outside" when full transparency (i.e., a code audit) is not feasible. In the case of commercial APIs, input and processing are opaque, thus, we can only manipulate the inputs in order to study the resulting outputs. In light of these challenges, as researchers, we aim to provide tools for auditing the output of CogS, in order to help developers and researchers better understand the benefits and risks of working with CogS, and to make the best choices of CogS to use according to their needs.

Social B(eye)as Dataset

Currently, we introduce the Social B(eye)as Dataset (SBD),[8] a collection of tags assigned to standardized people images by eight groups of taggers: six proprietary image tagging algorithms (referred to here as APIs), and "human taggers" - crowdworkers from two anglophone countries. The tags were all collected in October 2018. This dataset can be used to audit the social behaviors of a popular class of CogS - image tagging algorithms - as well as to study the behaviors of crowdworkers who have been paid to analyze the content of the same set of images.

This paper details release 1.0 of the SBD, which includes the following:

- **Output 1:** Metadata on the images used, along with raw, unedited tags from eight groups of taggers, and where applicable, metadata on the crowdworkers.
- **Output 2:** The tokenized & spellchecked (processed) tags for each image, from the eight groups of taggers.
- **Output 3:** Python script for calculating and exporting the vectors of the relative frequency per cluster/dimension, where the inputs are the .csv files of each sheet in the .xls file from Output 2.
- **Output 4:** Folder of dictionaries based on our typology, with one file for each subcluster and the corresponding tags, and one reference file with the raw tags and the corresponding processed tag.

More detailed information on the content and structure of the dataset can be found in the Metadata section.

Methodology

To build the Social B(eye)as Dataset, we took a set of 597 standardized images, passed each onto six APIs and two groups of crowdworkers, and processed the resulting tags. This process, as seen in Figure 1, is detailed below.

[3]https://azure.microsoft.com/en-us/services/cognitive-services/face/

[4]https://www.gartner.com/smarterwithgartner/the-algorithm-economy-will-start-a-huge-wave-of-innovation/

[5]https://blogs.windows.com/buildingapps/2017/02/13/cognitive-services-apis-vision/

[6]https://standards.ieee.org/project/7003.html

[7]http://www.privacy-regulation.eu/en/article-4-definitions-GDPR

[8]https://doi.org/10.7910/DVN/APZKSS

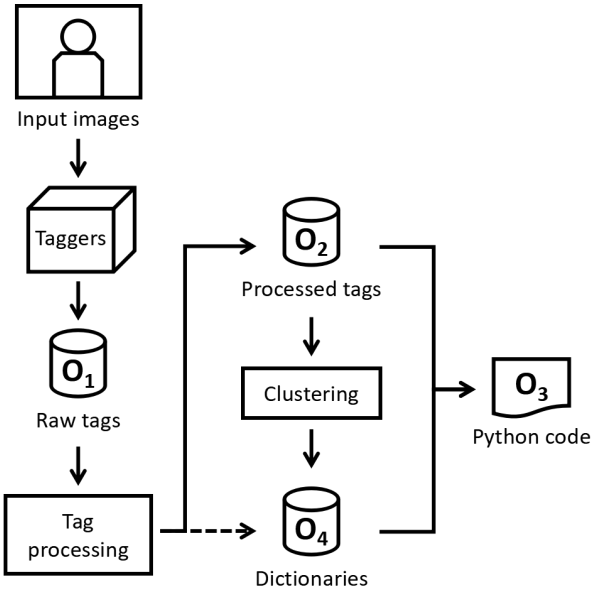

Figure 1: The data collection process.

Input: Chicago Face Database

While most images used by an API in its final implementation (i.e., a product) will probably be images "from the wild," this also introduces many variables that can affect the way people in an image are perceived, both by humans and machines. For that reason, to understand differences in the way that people of different races and genders are tagged, we used a standardized set of people images.

The Chicago Face Database (v2.0.3) (Ma, Correll, and Wittenbrink 2015), developed by psychologists and "intended for use in scientific research" including that on stereotyping and prejudice, has "high-resolution, standardized photographs of male and female faces of varying ethnicity between the ages of 17-65." [9] Each model is wearing the same gray t-shirt, standing in front of the same white background, looking straight at the camera, and the image was digitally edited to ensure standardization. Metadata for each image includes self-reported race, gender, and norming data (physical attributes such as face size, and subjective ratings such as attractiveness).[10]

We used the 597 portraits from the CFD with neutral expressions as our inputs. The distribution of the depicted persons' gender and race, self-reported from two and four mutually-exclusive categories respectively, is detailed in Table 1. Information about the depicted person's race, gender, and approximated age are included in Output 1, along with the image identifier which corresponds to the original "image code" in the CFD. In other words, those using the SBD can easily access the original images, by requesting access to the CFD, which is freely available.[11]

	Asian	Black	Latino/a	White	Total
Women	57	104	56	90	307
Men	52	93	52	93	290
Total	109	197	108	183	597

Table 1: Number of images by person's race and gender.

Taggers

To create the Social B(eye)as Dataset, we presented these 597 images to eight groups of taggers. As previously mentioned, six of these taggers were proprietary image tagging algorithms (APIs) while the other two groups were crowdworkers from two different countries (human analysts).

All of the six APIs are described by their providers as being general Machine Learning or Artificial Intelligence services. We decided to include only those using pre-trained models, representing a collection of tools that can be easily used by any developer, without any previous knowledge of machine learning. The six image tagging APIs used to process each image were:

- Amazon Rekognition Image[12] (hereon: Amazon)
- Clarifai[13] (hereon: Clarifai)
- Google Cloud Vision[14] (hereon: Google)
- Imagga Auto-tagging[15] (hereon: Imagga)
- IBM Watson Visual Recognition[16] (hereon: Watson)
- Microsoft Computer Vision[17] (hereon: Microsoft)

The crowdworking tasks were deployed in two anglophone countries to minimize meaning lost in translation, at the same time allowing for comparison between different cultures. The specific countries, India and US, were preferred as they both have a large population of available workers on the Figure Eight crowdsourcing platform.[18]

Tagging by image analysis APIs. We first passed the images through the six APIs. Specifically, we executed a series of RESTful calls in order to upload the CFD images into the CogS of the six different providers using HTTP Requests and saved their response as the output of this process.

Because the six APIs use different formatting for their output and do not follow the same structural guidelines, we did some pre-processing on the data to get the raw tags in a similarly-formatted output. For example, we extracted only the data that were useful for our research, excluding the

[9] https://chicagofaces.org/default/

[10] Subjective ratings, as well as estimated age, are reported in the CFD, based on assessments by at least 30 independent raters. See (Ma, Correll, and Wittenbrink 2015) for details.

[11] https://chicagofaces.org/default/

[12] https://aws.amazon.com/rekognition/image-features/

[13] https://clarifai.com/developer/guide/

[14] https://cloud.google.com/products/ai/

[15] https://imagga.com/solutions/auto-tagging.html

[16] https://www.ibm.com/watson/services/visual-recognition/

[17] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

[18] https://www.figure-eight.com/

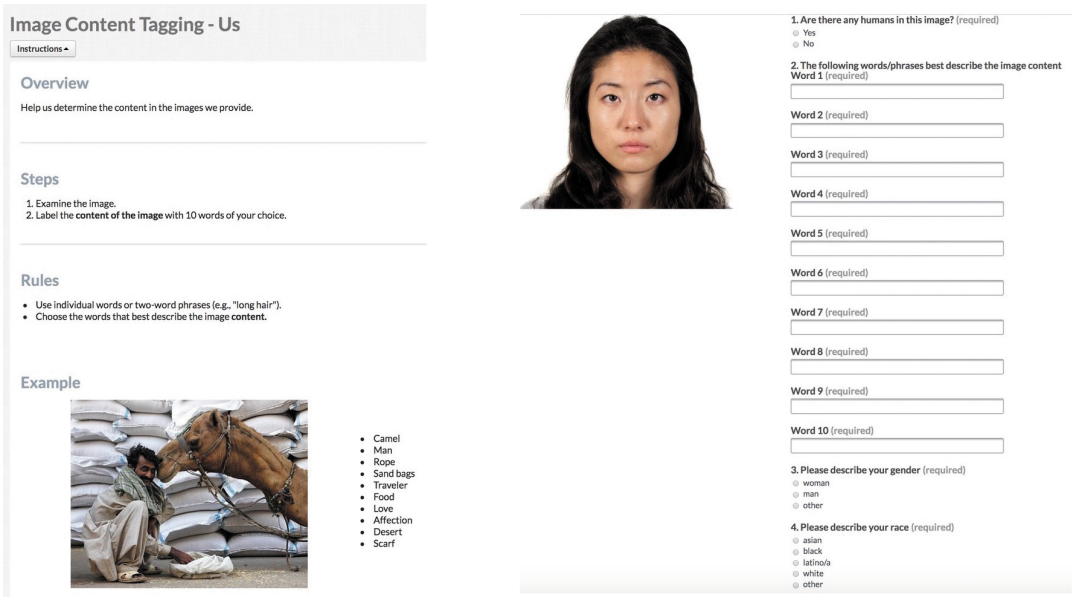

Figure 2: Task directions (left) and task interface (right).

rest (e.g., facial measurements, suggested captions, or additional inferences which some APIs provide). These otherwise unedited tags constitute Output 1, as a record per image for each API.

Tagging by Crowdworkers. We then used the same images to set up two tasks on Figure Eight, one each targeting workers in India and the U.S. As shown in the left side of Figure 2, the instructions were carefully modelled after the descriptions of the image tagging APIs, asking workers to "help us determine the content in the images," by providing "individual words or two-word phrases" that "best describe the image content."

In the example of an annotated image, which depicted a man and a camel, we provided 10 responses, as workers were required to do. Some of these were very concrete descriptions of the image content (e.g., man, rope, sand bags) while some were more abstract and interpretive (e.g., love, traveler). The form that workers used to complete the task is provided in Figure 2 (right side). They were first asked a very simple question that served as an attention check ("Are there any humans in the image?") They were then asked to provide 10 words or phrases to describe the image. Finally, they were asked to provide their own gender and race, where the choices for race corresponded to those used in the CFD. However, for both gender and race, workers could also respond with "other."

In both cases, we collected three responses per image from unique workers; an individual worker could describe up to 20 of our images. Workers in India were paid 20 cents per image, while workers in the U.S. received 30 cents per image. As the task took no longer than 120 seconds, this corresponds to an hourly wage of 6 and 9 USD, respectively. Workers were satisfied with the job, both in terms of the appropriateness of the set-up as well as the pay; in Figure-Eights' Contributor Satisfaction survey, our India task received a rating of 4.7 out of 5 (n=27 respondents), while the U.S. task was rated 4.9 out of 5 (n=28 respondents).

	India	**U.S.**
Unique workers	107	116
Median time on task	120 seconds	120 seconds
Maximum time on task	27 minutes	29 minutes

Table 2: Summary statistics for crowdwork.

We could not use test questions as a quality control mechanism, due to the open-ended nature of the task. However, to ensure quality, we enforced a minimum time per page (i.e., image) of 40 seconds. In addition, we used Figure-Eight's validators, using regular expressions to ensure that one-to-two words were provided. However, on occasion, workers submitted nonsense responses. Finally, we reviewed workers' responses and re-ran any observations that yielded less than four tags that were logical. In total, 88 (5%) of the responses in the India data were found to be invalid and were re-submitted for work.

Under U.S. and India crowdwork sheets in the Social B(eye)as Dataset, there are at least three responses per image per country, giving each image a minimum of six rows of human-produced tags.

Processing the raw tags

We tokenized and manually corrected/standardized the spelling of all unique tags from the eight groups of taggers. The tags by crowdworkers were often misspelled in ways that made the intention obvious - e.g. "eeys" would be corrected to "eyes". However, in cases where two different spellings were equally possible ("chik" could be "chick" or "cheek"), or where there were no obvious corrections ("hor-

Figure 3: Example of an image from the CFD (AF-248)

licks"), we left the raw tag as it was initially spelled. Further, where a tagger (human or machine) used a tag with more than one word, we substituted the spaces (" ") with underscores ("_") to handle the phrase as one tag.

Output 4 has a .csv file where the first value in each row is the raw tag with the original spelling, and the second value is the processed tag with the corrected spelling. Output 2 (processed tags) is a version of Output 1 (raw tags) where the raw tag has been replaced with the corresponding "corrected" processed tag.

(Re)Defining thematic clusters & Categorizing tags

We aimed to create a typology that maps the tags to a set of common concepts. Given that taggers use different vocabularies (i.e., might describe the same underlying concept in different ways), the typology helps us compare how taggers perceive and describe the people images. We applied an inductive thematic analysis (Herring 2009) to the tag sets, as described below. Manual clustering was preferred over automatic clustering (e.g. via sentiment analysis algorithms) as our aim was to look for embedded human biases, and such algorithms may come with biases which have not been discovered yet. In contrast, manual clustering with the same few researchers minimizes the biases in the process.

We first started by clustering the tags from the APIs, since the tags were fewer and easier to understand. We treated all tags as stand-alone; each tag was judged on whether the meaning would have been clear if we didn't know the inputs were all images with one person, and where there was a tag with synonyms, we did not consult the accompanying tags from the same tagger to discern the specific meaning.

Using these initial clusters that emerged as guidance, we started categorizing the tags from the crowdworkers as well. Once the initial categorizing was finished, it became clear that we needed to redefine existing clusters and add new ones. With this new (and final) typology, three researchers each went through the entire set of unique tags from the eight groups of taggers. Their results were compared, the disagreements discussed, and the contested concepts recategorized through an inductive, iterative process.

There were 21 distinct categories (subclusters) that resulted from this clustering, which can be further grouped into five "superclusters" of common themes. The hierarchy of the clusters, examples of tags in each cluster, and further information can be found in Table 4. The subclusters, along with the tags that are categorized within, are presented as separate .csv files in Output 4.

Tagger	Processed Tags
Amazon	human, people, person, face, freckle, portrait, female, woman
Clarifai	woman, cute, portrait, one, pretty, people, funny, girl, look, fashion, eye, face, facial_expression, adult, isolated, looking, friendly, lips, wear, young
Google	face, eyebrow, cheek, chin, nose, forehead, head, beauty, neck, lip
Imagga	mug_shot, photograph, representation, creation, portrait, face, model, attractive, pretty, hair, adult, smile, eyes, fashion, person, sexy
Microsoft	person, smiling, posing, woman, wearing, front, shirt, black, blue, young, photo, white, holding, hair, donut, man, large, standing
Watson	person, anchorperson, official, coal_black_color, light_brown_color
Indian Crowdworkers	woman, young, wheatish_complexion, beautiful, straight_hair, long_hair, thick_lips, thin_eyebrows, oval_face, asian_descent
US Crowdworkers	pretty_girl, brown_skin, black_hair, brown_eyes, smooth_skin, makeup_free, jewelry_free, pleasant_face, full_lips, grey_sweatshirt

Table 3: Tags collected for the example image in Figure 3.

Creating vectors

In order to compare the behaviors of the eight taggers in a straight-forward manner, we "scored" each image description from a tagger with respect to our typology. In other words, we represented each image description as a vector in a 25-dimensional space, corresponding to the description's reference to each of the 25 concept clusters (including both super- and subclusters). Specifically, we record the relative frequency per cluster/dimension (i.e., the proportion of tags in a description that map onto a given concept cluster).

For this purpose, we created a python script which calculates and exports these vectors, using the .csv files of each sheet in Output 2. Output 2 is an .xls file that consists of the image identifier and the corresponding processed tags from each tagger, on separate sheets.

In addition to the above representation, we also represented each image/tagger description with respect to the sub- and superclusters separately. In other words, we produced the 20-dimensional and 5-dimensional vector representations as well. The relevant python script can be found under Output 3, along with a set of dictionaries that the script uses for clustering (i.e., scoring) purposes under Output 4.

Cluster		Description	Examples
Demographics		*Tags that describe the inferred gender, age and/or origin(s) of the depicted person*	
	Masculine	Tags that refer to a masculine gender identity or expression	son, masculinity, him
	Feminine	Tags that refer to a feminine gender identity or expression	woman, girl, latina
	Nonbinary	Tags that refer to an aspect of gender other than masculine and feminine	androgynous, gender, transgender
	Age	Tags that refer to the person's age	millenial, girl, thirties
	Race	Tags that refer to the person's race, ethnicity, nationality, or religion	nigerian, migrant, light_skin
Concrete		*Tags that describe directly observable attributes of the image or the depicted person*	
	Actions	Tags that refer to the movement of the body or face	standing, squint, smiling
	Body	Tags that refer to the body, a feature exclusive to the body, or the species	big_nose, scars, human
	Hair	Tags that refer to the hair of the person, including facial hair	blond, shaven, weird_hair
	Clothing	Tags that refer to clothing, accessories, and makeup	gray_tshirt, wig, lined_eyes
	Colors	Tags that refer to colors	red, dark_roots, pigmented
	Meta	Tags that refer to the image itself, including location, type, and purpose	indoors, portrait, passport_photo
	Shape	Tags that refer to the shape, size/amount, or position of the person or something about the person	crooked_nose, fat, nonsymmetrical
Abstract		*Tags that describe the inferred, subjective or conceptual attributes of the person*	
	Judgement	Tags that describe an opinion or subjective description	normal, beautiful, photogenic
	Traits	Tags that refer to a personality trait or enduring characteristic	extrovert, stubborn, macho
	Emotion	Tags that refer to an emotional, mental, or temporary physical state	happy, concentrated, ill
	Occupation	Tags that refer to a job, a field of work, or a social role	athlete, son, damsel
Inflammatory		Tags that are unambiguously racist and/or sexist	
Other		*Tags that do not fit into any of the previous clusters because their meaning is not clear, tags that refer to an outlier concept, and tags that refer to the absence of a concept*	
	Ambiguous	Tags that are understandable but could refer to one of two or more meanings (tags are not included in the potential clusters)	blue, cold, let_down
	Inconclusive	Tags that are not understandable, including tags in a language other than English	horlicks, grand_nez, nise
	Lack	Tags that negate a concept, referring to the lack of something (these tags are also included in the cluster that houses the concept they are negating)	beardless, not_shaven, absent_ears
	Misc	Tags that are understandable with a singular meaning, but do not fit into any of the other clusters	pizza, welfare, winter

Table 4: Thematic cluster names, explanations, & example tags.

Typology: cluster descriptions

Here, we describe in more detail the meaning of each thematic cluster in our typology. As mentioned, tags similar in meaning were brought together to form "subclusters," which are grouped into five "superclusters." The hierarchy, names, descriptions, and example tags of each cluster (both super- and subclusters) can be seen in Table 4 and in ReadMe files in the dataset.

It's important to note that the subcluster names were simplified for convenience in referencing them; however, they refer to concepts larger than the title. For example, the "Race" subcluster is not limited to race, but also includes adjacent concepts such as ethnicity, nationality, and religion as well. The clusters are not mutually-exclusive; where a tag (e.g. "girl") implies more than one concept ("age" and "feminine"), the tag appears in all applicable clusters.

Metadata

The Social B(eye)as Dataset consists of four distinct outputs (as described in the Introduction). Here, we describe the format, structure, and content of each file within the dataset.

Output 1: Raw tags

1 .xls file with 8 sheets (1 per tagger group)

The first four columns in every sheet corresponds to: the image identifier (e.g. AF-248, the same code that is used to identify the image in the Chicago Face Database), the race, gender, and the approximated age of the depicted person. (See subsection on Input: CFD for more detail as well as the appropriate reference.) The next two columns of the two crowdworker sheets correspond to the race and gender of the crowdworker completing the task (these columns are not present in the six sheets for the APIs). Each of the following columns - for all sheets - correspond to one raw (i.e., unprocessed, untokenized) tag. The number of tags varies between the eight groups of taggers, sometimes also varying within the tagger group.

The six sheets for the APIs have 598 rows (title row + one row per image), while the two sheets for the crowdworkers have 1792 (US) and 1797 (India) rows (title row + minimum of three rows/responses per image).

Output 2: Processed tags

1 .xls file with 8 sheets (1 per tagger group)

The first column in every sheet corresponds to the image identifier, same as Output 1. Each of the following columns, for all sheets, corresponds to one processed tag.

This file was created by taking the raw tags per image from Output 1 and replacing each raw tag with the corre-

Cluster	Amazon	Clarifai	Google	Imagga	Microsoft	Watson	India	US	All Taggers
Demographics	8	14	7	10	8	18	243	164	354
Masculine	1	5	3	4	3	6	47	40	70
Feminine	4	2	1	1	3	7	70	37	91
Nonbinary	0	0	0	0	0	0	2	2	4
Age	6	10	6	7	6	16	103	67	153
Race	0	1	0	0	0	0	135	86	184
Concrete	23	38	38	31	47	48	1053	1097	1879
Action	1	7	3	5	10	0	72	89	147
Body	9	7	19	6	7	12	606	617	1014
Hair	6	9	13	5	1	10	216	233	393
Clothing	5	3	1	8	11	11	66	74	150
Color	1	3	4	2	9	14	200	173	332
Meta	2	8	2	6	7	4	30	21	57
Shape	0	4	1	0	2	0	401	425	679
Abstract	0	38	3	10	0	14	412	345	677
Judgement	0	7	1	5	0	0	158	100	229
Traits	0	20	0	1	0	0	127	98	200
Emotion	0	8	2	2	0	0	129	146	228
Occupation	0	5	0	2	0	14	17	11	44
Inflammatory	0	0	0	0	0	0	4	5	7
Other	10	14	19	9	21	12	686	697	478
Ambiguous	9	7	19	6	7	12	606	617	147
Inconclusive	0	0	0	0	0	0	33	29	62
Lack	0	1	0	0	0	0	43	28	65
Misc	1	6	0	3	14	0	53	59	125
Total Unique	32	95	48	54	71	75	1626	1567	2823

Table 5: Number of *unique* tags used by the taggers per thematic cluster.

sponding processed tag (the pairing of which can be found in a file in Output 4). Therefore, the sheets within this file have the same number of rows as the sheets in Output 1.

Output 3: Python script

1 .py file

As mentioned before, this script calculates the vectors of the relative frequency per cluster/dimension (i.e., the proportion of tags in a description that map onto a given concept cluster).

The script takes a .csv file (Output 2) that represents the list of processed tags given for each target as input and outputs the vectors in a .csv format where the first column is the image identifier and the rest are one column for each super- and subcluster value.

Output 4: Dictionaries

22 .csv files

There are 20 .csv files which correspond to the subclusters and one .csv file that corresponds to the "Inflammatory" supercluster (which does not contain any subclusters). Together, they contain all unique tags in the dataset, categorized within our typology as described in the previous section. The file name of these 21 .csv files corresponds to the title of the cluster (e.g. age cluster: *age.csv*), where each row has one of the unique processed (tokenized & spellchecked) tags categorized within that cluster.

The remaining (1) file (*corrections_dict.csv*) is such that on each row, the first value is the raw tag (original output from taggers), followed by the second value which is the corresponding processed tag.

Using the Dataset

In this section, we first disclose some limitations that users of the SBD should recognize. Next, we present some initial findings based on our dataset, as well as some use cases for future applications of this rich dataset.

Limitations

With the large number of free-text inputs, we obtained many tags that were misspelled, some beyond comprehension. We corrected many of these, given that there was a consensus on what was intended. However, 143 tags were deemed unusable for analysis regarding meaning, and have been placed in the "Inconclusive" subcluster.

It must be stated that any manual clustering task, on a large set of inputs such as our tags, is subject to human error. In particular, it is possible that the researchers were influenced by "respondent fatigue," losing focus towards the tags at the bottom of each list relative to the tags at the top. Similarly, contested concepts may have been described through the use of different word-tags (e.g., smiling, frowning). A few of these tags, which are similar in meaning could have been missed, given the large number of tags analyzed.

That said, none of the above limitations apply to tags which have been used frequently, which constitutes the majority of tags in the dataset. Therefore, any human error on behalf of the researchers or the crowdworkers is minimal,

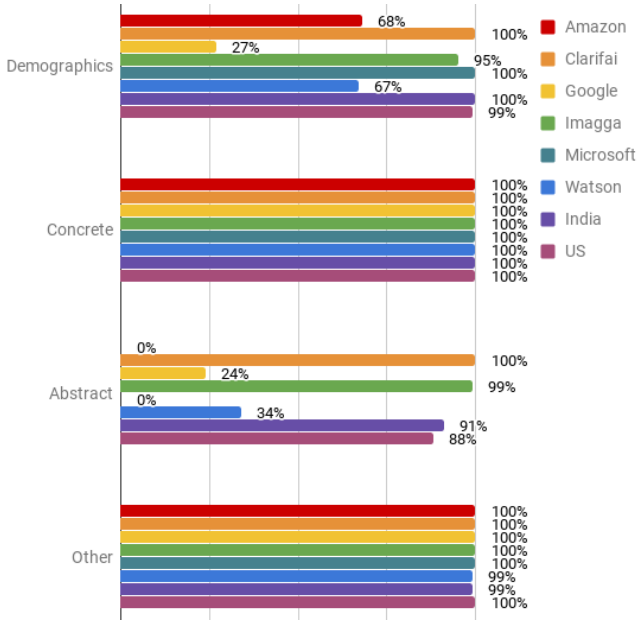

Figure 4: Proportion of images with at least one tag in each supercluster, by tagger.

corresponding to a very small percentage of the total number of tags and thus, is not expected to affect the correlations and conclusions drawn. More importantly, the manual clustering, as stated earlier in the Methodology section, is preferable over automatic clustering as the aim is to look for embedded biases within the technology we are auditing.

Basic findings

Table 5 shows the number of unique tags used by each tagger, with respect to cluster. A closer look shows that, as opposed to the diverse vocabulary from the crowdworkers, there are categories for which there are (almost) no tags from the APIs. For example, there is only one tag from the APIs that falls into the Race and Lack categories, and no tags at all which are Inconclusive, Inflammatory, or Nonbinary.

Figure 4 shows the percentage of images which have a tag in each of the superclusters. We can easily see that every single tagger has commented on the concrete (i.e. observable) attributes of the image. It's interesting to note that two APIs (Amazon and Microsoft) have not used any Abstract (i.e. intangible concept) tags while, one API (Clarifai) has used at least one Abstract tag on every image.

Looking at the Demographics (Figure 5) and Abstract (Figure 6) superclusters in more detail, we can see that some taggers show large differences in how men and women are tagged. For example, the Google image tagging API appears to be ten times more likely to tag images of men with a Demographics tag, while Amazon is much more likely to tag women than men. More robust analysis and its implications can be found in (Kyriakou et al. 2019).

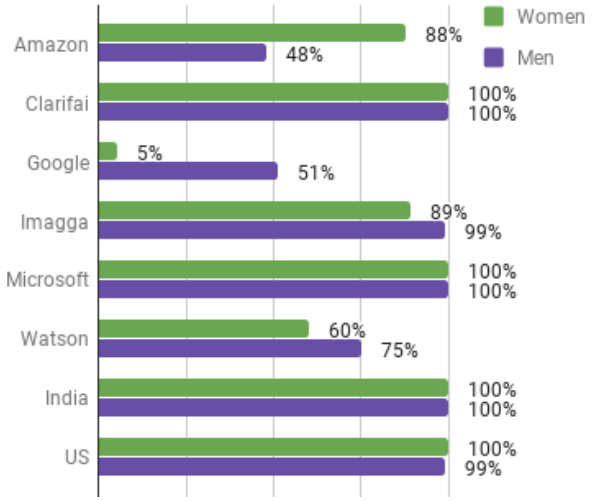

Figure 5: Proportion of images with at least one tag in the **Demographics** supercluster, by gender of depicted person.

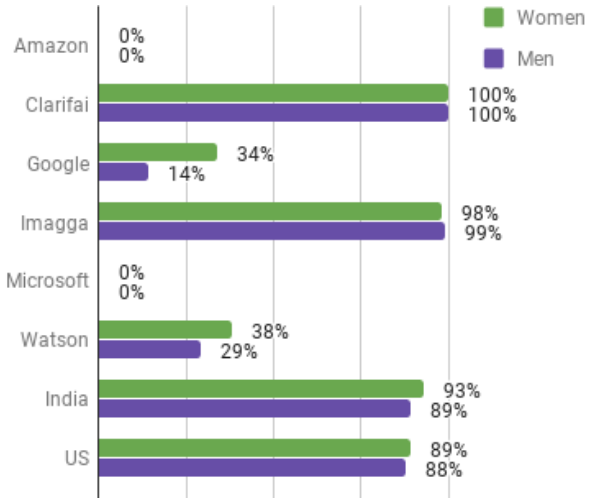

Figure 6: Proportion of images with at least one tag in the **Abstract** supercluster, by gender of depicted person.

Further analysis & Use cases

Much can be said with regard to the proportion of tags used on different groups of people, the number of tags in use by the APIs (that we can see so far), the differences in what taggers choose to comment on, and more. Further analysis can show whether there are correlations between any of the different dimensions we detail in this dataset.

The Social B(eye)as Dataset and its methodology can be used to audit other commercial image tagging CogS/APIs for bias regarding gender and/or race. But even beyond computer vision audits and the social biases of algorithmic processes, this dataset can also be used for research in many fields. Some topics include:

- Cultural differences in crowdwork
- Quality issues in crowdwork

- Human bias regarding gender, race, and/or age
- People's perceptions of the tags produced by algorithms
- Language and concept diversity in people perception
- Correlations between face shape and people perception
- Linguistic similarities and differences between cultures and/or Cognitive Services.

As an example, in (Barlas et al. 2019), we looked into whether crowdworkers found the tags produced by one of the APIs or those produced by other crowdworkers to be more "fair," and how they would describe "fairness."

FAIR Data

In this section, we explain how we have made the data Findable, Accessible and Interoperable, in order to increase data Re-use (FAIR). First, the dataset is freely accessible through Dataverse,[19] with the following citation:

Barlas, Pinar; Kyriakou, Kyriakos; Kleanthous, Styliani; Otterbacher, Jahna, 2019, "Social B(eye)as Dataset", https://doi.org/10.7910/DVN/APZKSS, Harvard Dataverse, V1.

The format of the files consists of Comma Separated Values (CSV) format and eXceL Spreadsheet (XLS) format, so that the data can be handled with any application or script, exported in other formats and re-used for other purposes. This enables the interoperability of our dataset and increases the data re-use.

Acknowledgments

This project is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 739578 (RISE), 810105 (CyCAT) and the Government of the Republic of Cyprus (RISE).

References

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2019. What Makes an Image Tagger Fair? Proprietary Auto-tagging and Interpretations on People Images. In *Proceedings of the 27th ACM Conference On User Modelling, Adaptation And Personalization*, UMAP '19. ACM.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 77–91.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1):2053951715622512.

Deeb-Swihart, J.; Polack, C.; Gilbert, E.; and Essa, I. A. 2017. Selfie-presentation in everyday life: A large-scale characterization of selfie contexts on instagram. In *ICWSM*, 42–51.

Garimella, V. R. K.; Alfayad, A.; and Weber, I. 2016. Social media image analysis for public health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5543–5547. ACM.

Herring, S. C. 2009. Web content analysis: Expanding the paradigm. In *International handbook of Internet research*. Springer. 233–249.

Hu, Y.; Manikonda, L.; Kambhampati, S.; et al. 2014. What we instagram: A first analysis of instagram photo content and user types. In *ICWSM*.

Kocabey, E.; Ofli, F.; Marin, J.; Torralba, A.; and Weber, I. 2018. Using computer vision to study the effects of bmi on online popularity and weight-based homophily. In *International Conference on Social Informatics*, 129–138. Springer.

Kyriakou, K.; Barlas, P.; Kleanthous, S.; and Otterbacher, J. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*, ICWSM-19. AAAI.

Liu, L.; Preotiuc-Pietro, D.; Samani, Z. R.; Moghaddam, M. E.; and Ungar, L. H. 2016. Analyzing personality through social media profile picture choice. In *ICWSM*, 211–220.

Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47(4):1122–1135.

Rhue, L. 2018. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765*.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 1–23.

[19] http://dataverse.harvard.edu